Week 2

Case Study – Understanding the data

- Descriptive statistics
  - Univariate statistics
  - Pairwise correlations
  - Histogram

$$f(a) + \sum_{k=1}^{n} \frac{1}{k!} \frac{d^k}{dt^k}\bigg|_{t=0} f(u(t)) + \int_0^1 \frac{(1-t)^n}{n!} \frac{d^{n+1}}{dt^{n+1}} f(u(t))\, dt.$$

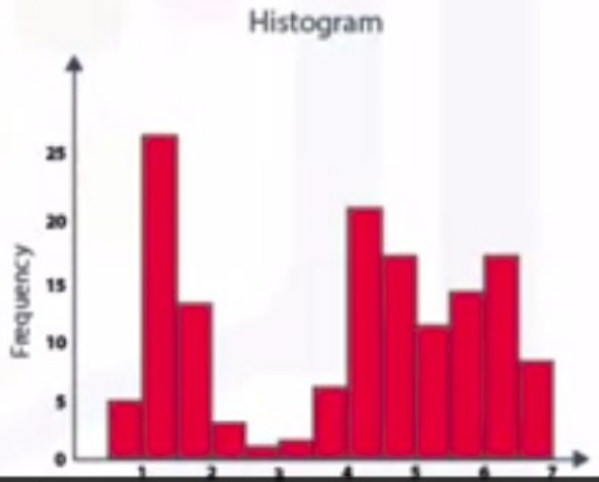
$F_{X,Y}(x, y)$ satisfies

$$F_{X,Y}(x, y) = F_X(x) F_Y(y),$$

or equivalently, their joint density $f_{X,Y}(x, y)$ satisfies

$$f_{X,Y}(x, y) = f_X(x) f_Y(y).$$

Histograms are a good way to understand how values or a variable are distributed, and what sorts of data preparation may be needed to make the variable more useful in a model.



First, these statistics included Hearst, univariates, and statistics on each variable, such as mean,

Case study – Looking at data quality

- Data quality
  - Missing values
  - Invalid or misleading values

From the information provided, certain values can be re-coded or perhaps even dropped if

Question

The Data Understanding stage encompasses sorting the data.

- ○ True.
- ◉ False.

Correct
Correct.

Skip Continue

Examples of data cleansing

	A	B	C	D	E
1	Name	Date	Age	Location	Country
2	John Doe	2012 02 20	32	ON	CAN
3	May Lag	2013 02 33	2	ON	CA
4	Henry Oon	30-Sep-12	35	Ontario	CANADA
5	Kelly, Tom	2015 02 20	65	ON	CA
6	John Kell	2016 02 20		AB	CA
7	Henry Oon	30-Sep-12	35	Ontario	CANADA
8					

Invalid Values

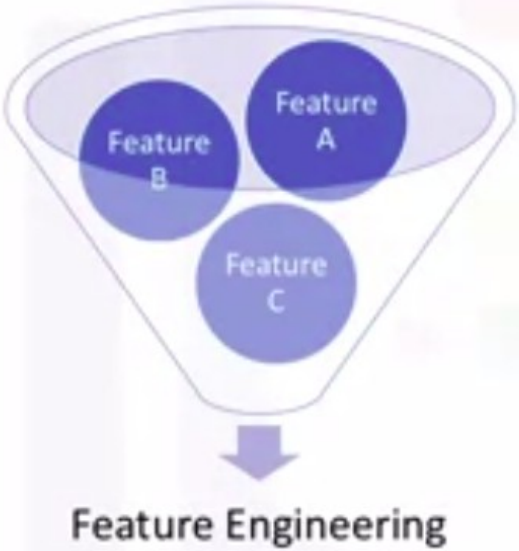
Missing Data

Remove Duplicates

Formatting

or invalid values and removes duplicates, toward ensuring that everything is properly

Using domain knowledge



Feature engineering is the process of using domain knowledge of the data to create features that make the machine learning algorithms work.

Feature engineering is critical when machine learning tools are being applied to analyze the data.

Feature engineering is critical when machine learning tools are being applied to analyze

Congratulations! You passed!

Grade received 100% **Latest Submission Grade** 100% **To pass** 66% or higher

Go to next item

1. In the case study, working through the **Data Preparation** stage, it was revealed that the initial definition was not capturing all of the congestive heart failure admissions that were expected, based on clinical experience.

1 / 1 point

- ○ True
- ◉ False

Correct
Correct. It was working through the **Data Understanding** staging that the initial definition was found to be incomplete.

2. Select the correct statement about what data scientists do during the Data Preparation stage.

1 / 1 point

- ○ During the Data Preparation stage, data scientists define the variables to be used in the model.
- ○ During the Data Preparation stage, data scientists determine the timing of events.

2. Select the correct statement about what data scientists do during the Data Preparation stage.

1 / 1 point

- ○ During the Data Preparation stage, data scientists define the variables to be used in the model.
- ○ During the Data Preparation stage, data scientists determine the timing of events.
- ○ During the Data Preparation stage, data scientists aggregate the data and merge them from different sources.
- ○ During the Data Preparation stage, data scientists identify missing data.
- ◉ All of the above statements are correct.

Correct
Correct.

3. The Data Preparation stage is a very iterative and complicated stage that cannot be accelerated through automation.

1 / 1 point

- ○ True
- ◉ False

Correct
Correct.

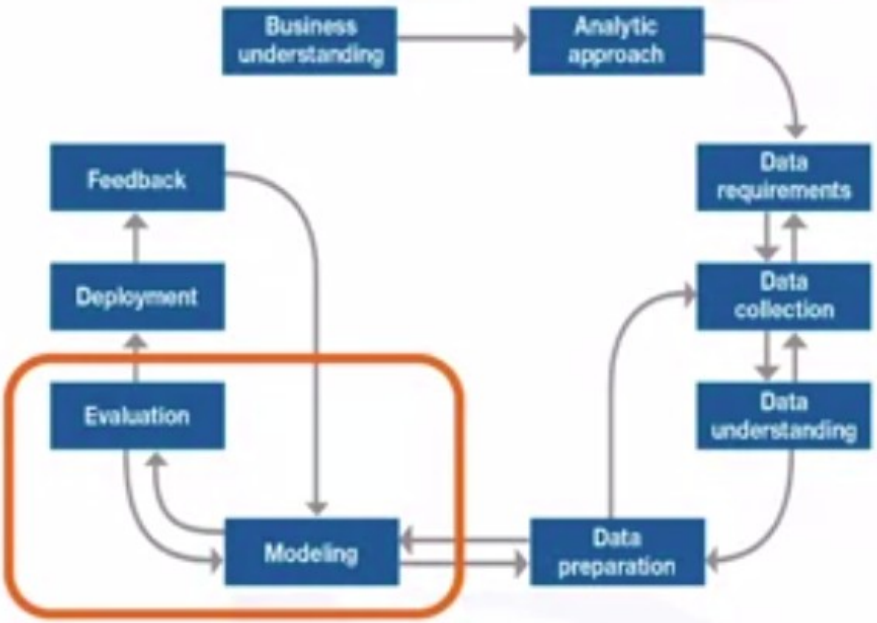
Lesson Summary

In this lesson, you have learned:

- The importance of descriptive statistics.
- How to manage missing, invalid, or misleading data.
- The need to clean data and sometimes transform it.
- The consequences of bad data for the model.
- Data understanding is iterative; you learn more about your data the more you study it.

Mark as completed

From Modeling to Evaluation



Modeling

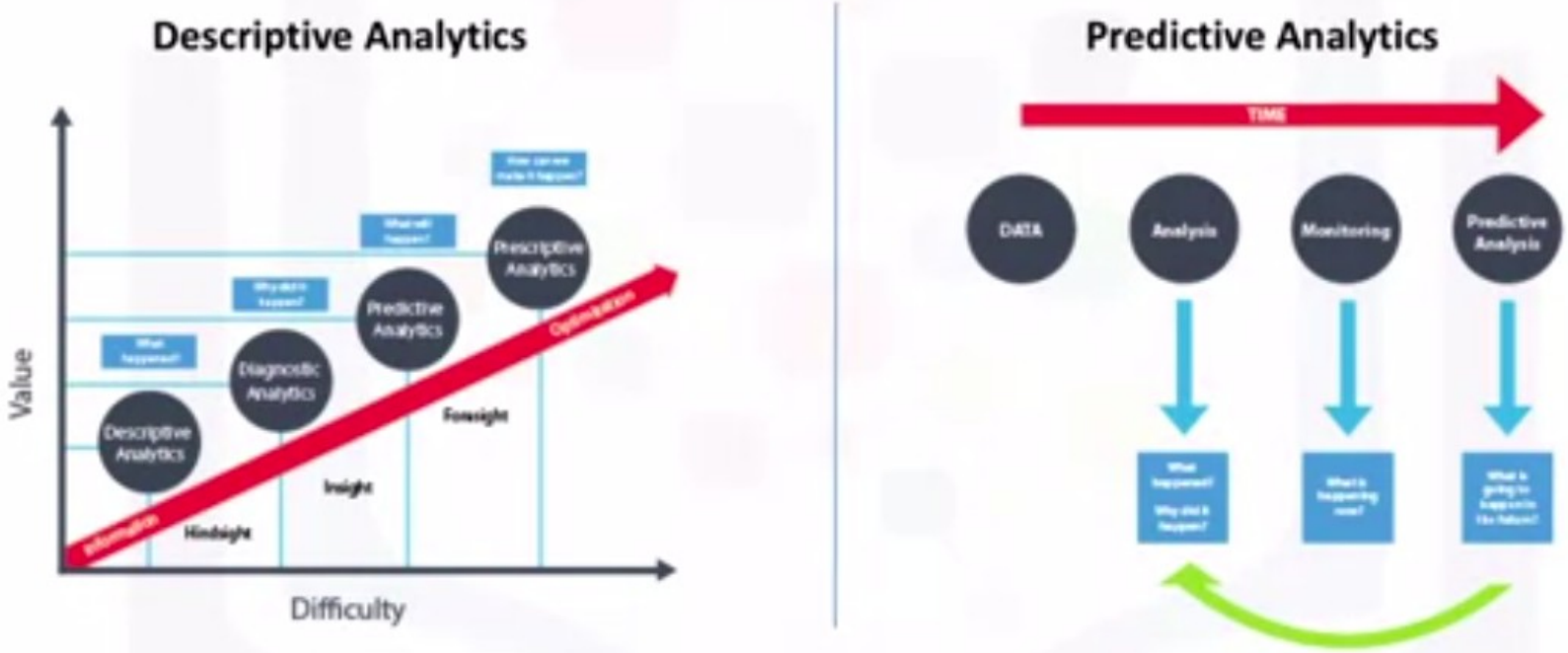
- *In what way can the data be visualized to get to the answer that is required?*

Evaluation

- *Does the model used really answer the initial question or does it need to be adjusted?*

This portion of the course is geared toward answering two key questions:

Data Modeling – Using Predictive or Descriptive?



An example of a descriptive model might examine things like: if a person did this,

Understanding the question

1. Understand the question at hand
2. Select an analytic approach or method to solve the problem
3. Obtain, understand, prepare, and model the data

of the problem at hand, and the appropriate analytical approach being taken.

Question

A training set is used to build a predictive model.

- (●) True.
- ( ) False.

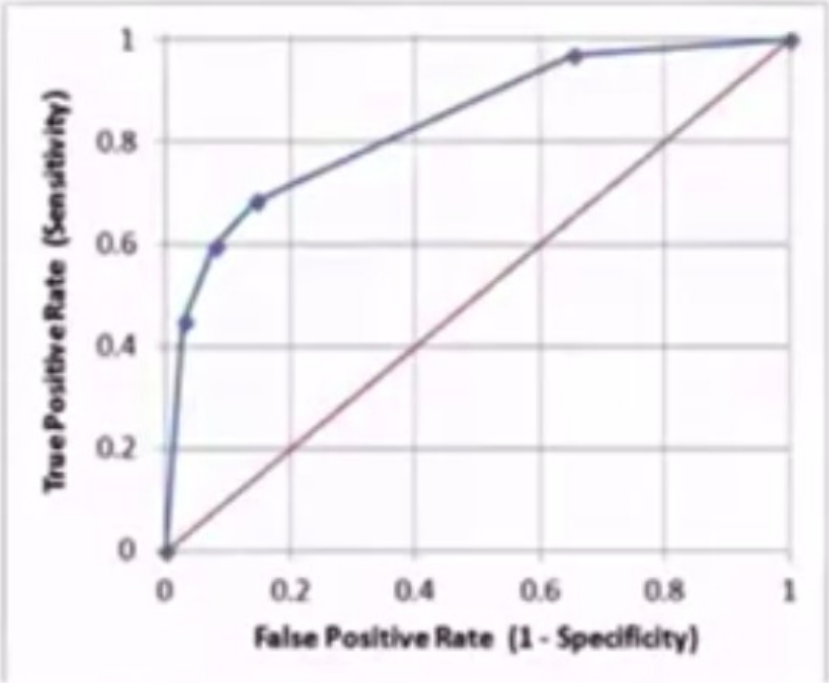
✓ **Correct**
Correct.

Skip | Continue

Case Study – Using the ROC curve

Diagnostic tool for classification model evaluation

- Classification model performance
- True-Positive Rate vs False-Positive Rate
- Optimal model at maximum separation



which was first developed during World War II to detect enemy aircraft on radar.

Question

Model evaluation can have two main phases: a diagnostic measures phase and statistical significance testing.

- False.
- True.

Correct
Correct.

Skip Continue

Lesson Summary

In this lesson, you have learned:

- The difference between descriptive and predictive models.
- The role of training sets and test sets.
- The importance of asking if the question has been answered.
- Why diagnostic measures tools are needed.
- The purpose of statistical significance tests.
- That modeling and evaluation are iterative processes.

Completed Go to next item

Week 3

Case Study – Understand the results

Assimilate knowledge for business

- Practical understanding of the meaning of model results
- Implications of model results for designing intervention actions

In preparation for solution deployment, the next step was to assimilate the knowledge

Case Study – Gathering application requirements

Application requirements

- Automated, near-real-time risk assessments of CHF inpatients
- Easy to use
- Automated data preparation and scoring
- Up-to-date risk assessment to help clinicians target high-risk patients

of congestive heart failure.

From Deployment to Feedback

Once the model is evaluated and the data scientist is confident it will work, it is deployed and put to the ultimate test

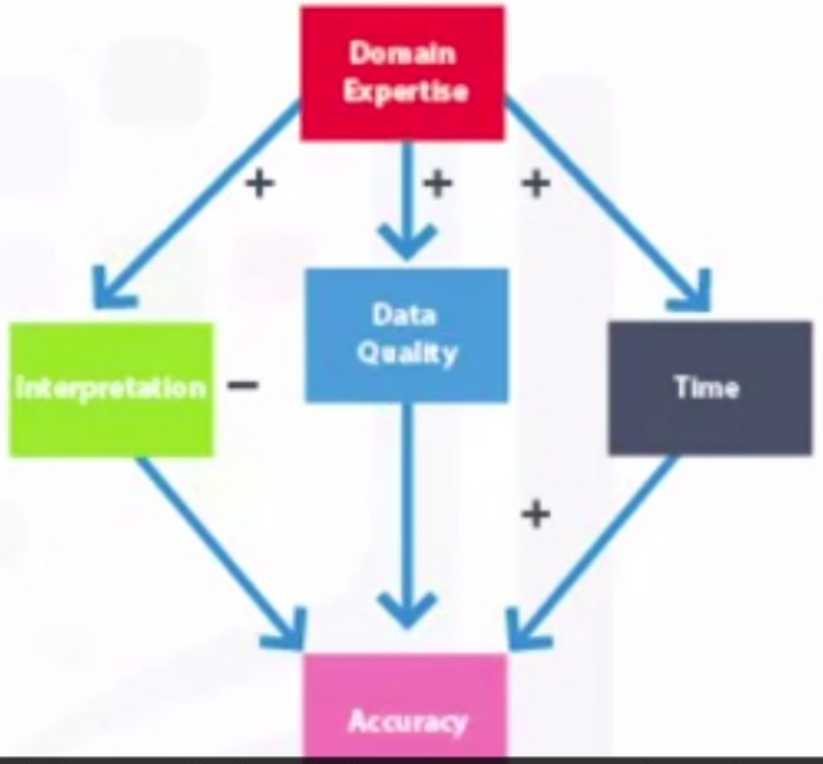
- Actual real-time use in the field

Once the model is evaluated and the data scientist is confident it'll work, it is deployed

Case Study – Assessing model performance

Define review process

- To measure results of applying the risk model to the CHF patient population
- Track patients who received intervention
  - Actual readmission outcomes
- Measure effectiveness of intervention
  - Compare readmission rates before & after model implementation

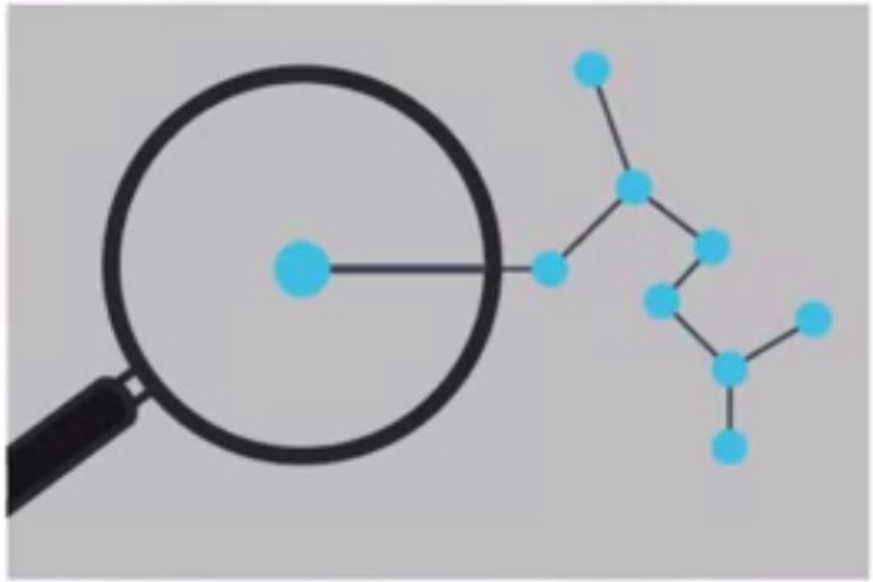


for measuring the results of a "flying to risk" model of the congestive heart failure

Case Study – Refinement

Refine model

- Initial review after the first year of implementation
- Based on feedback data and knowledge gained
- Participation in intervention program
- Possibly incorporate detailed pharmaceutical data originally deferred
- Other possible refinements as yet unknown

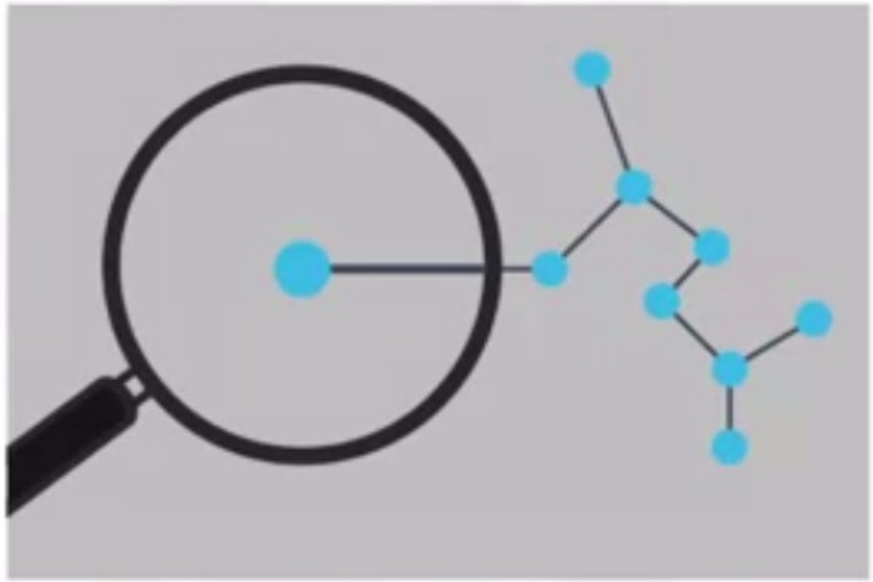


Other refinements included: Incorporating information about participation

Case Study – Refinement

Refine model

- Initial review after the first year of implementation
- Based on feedback data and knowledge gained
- Participation in intervention program
- Possibly incorporate detailed pharmaceutical data originally deferred
- Other possible refinements as yet unknown



during the feedback stage.

Question

The data science methodology is highly iterative, ensuring the refinement at each stage in the game.

- ○ False.
- ◉ True.

✓ **Correct**
Correct.

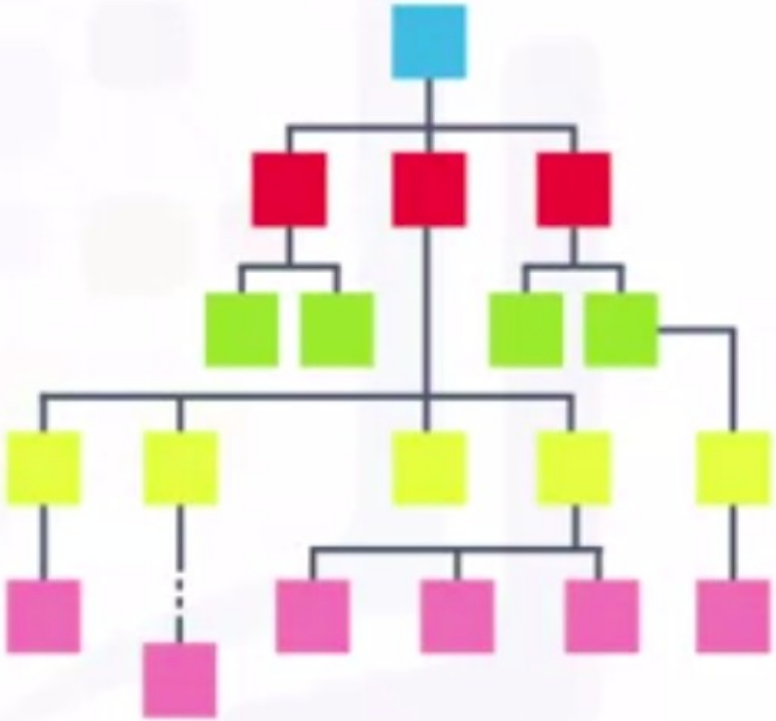
Skip | Continue

Case Study – Redeployment

Review and refine intervention actions

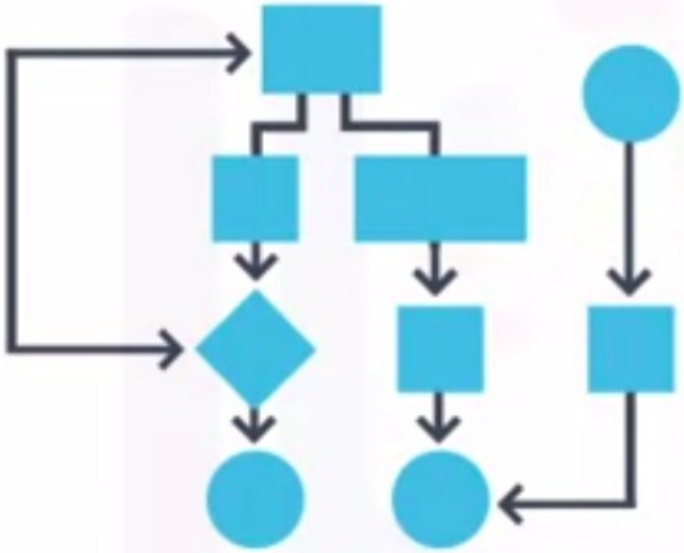
Redeploy

- Continue modeling, deployment, feedback, and refinement throughout the life of the intervention program



This is the end of the Feedback portion of this course.

From problem to approach



Thinking like a data scientist!

- ✓ Forming a concrete business or research problem,
- ✓ Collecting and analyzing data,
- ✓ Building a model, and
- ✓ Understanding the feedback after model deployment.

Learned importance of:

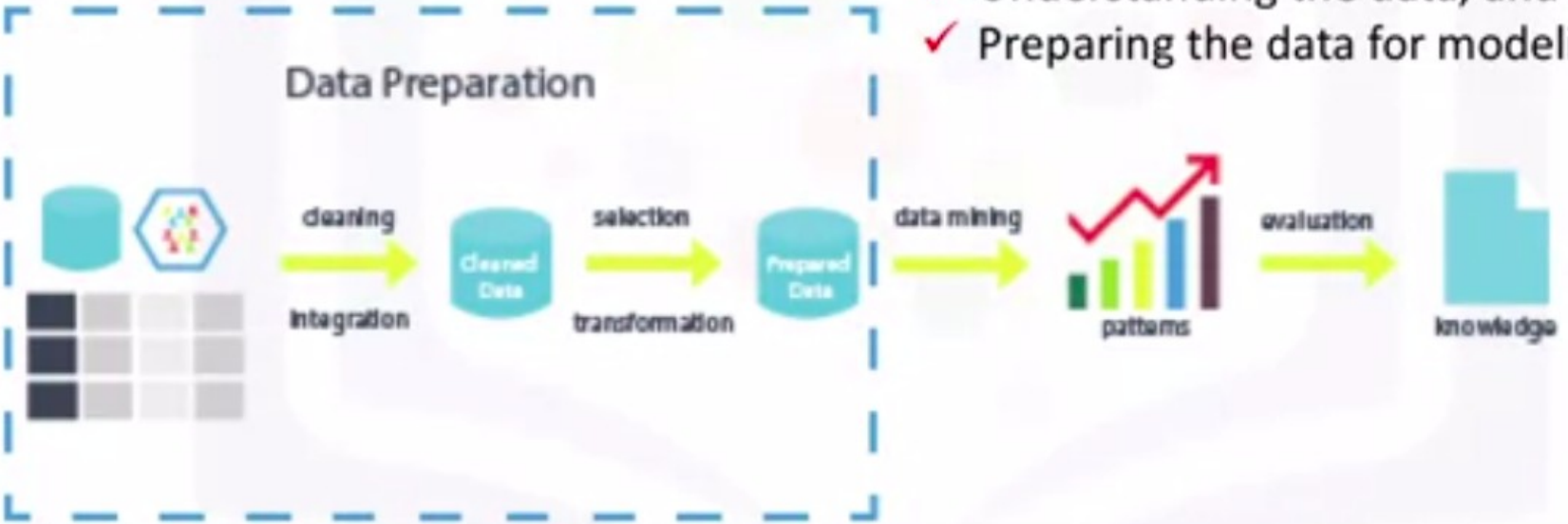
- ✓ Understanding the question, and
- ✓ Picking the most effective analytic approach

forming a concrete business or research problem, collecting and analyzing data,

To working with the data

Learned to work with data!

- ✓ Determining the data requirements,
- ✓ Collecting the appropriate data,
- ✓ Understanding the data, and then
- ✓ Preparing the data for modeling!

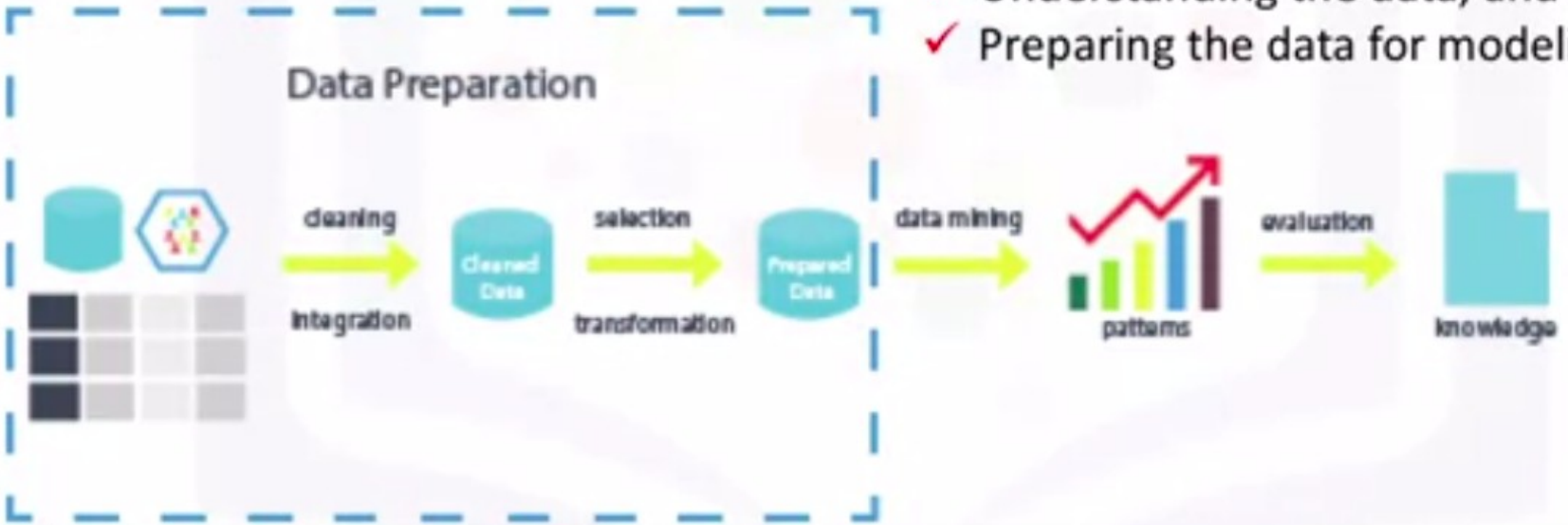


understanding the data, and then preparing the data for modeling!

To working with the data

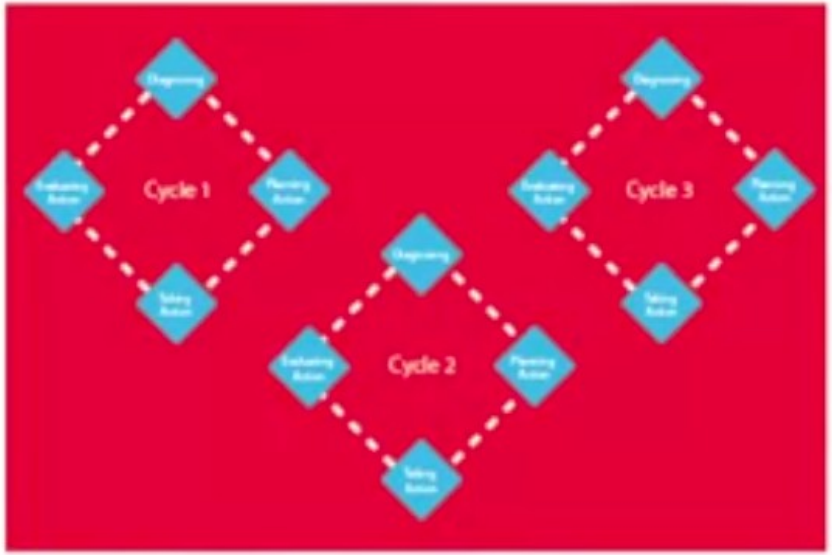
Learned to work with data!

- ✓ Determining the data requirements,
- ✓ Collecting the appropriate data,
- ✓ Understanding the data, and then
- ✓ Preparing the data for modeling!



understanding the data, and then preparing the data for modeling!

To deriving the answer



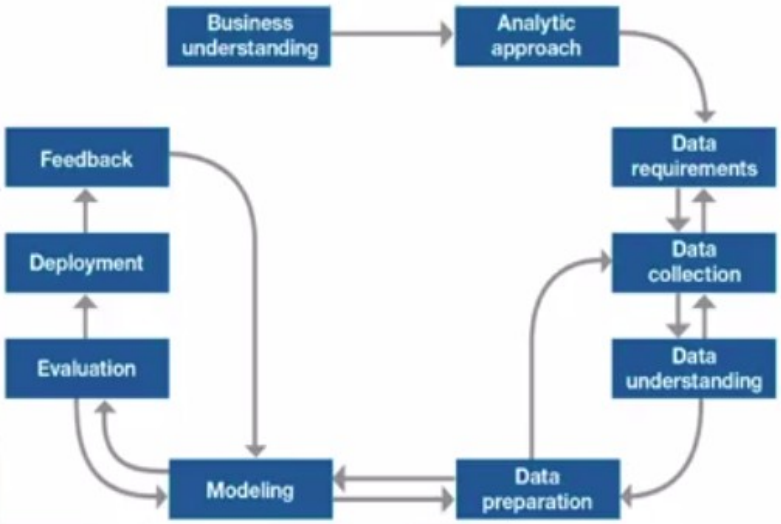
Once the analytic approach is selected, learned how to derive the answer:

- ✓ Evaluating and deploying the model,
- ✓ Getting feedback on it, and
- ✓ Using that feedback constructively so as to improve the model.

Remember that the stages of this methodology are ***iterative***!

You've also learned how to model the data by using the appropriate analytic approach,

The methodology in a nutshell…



The **Data Science Methodology** aims to answer the following 10 questions in this prescribed sequence:

From problem to approach:

1. *What is the problem that you are trying to solve?*
2. *How can you use data to answer the question?*

Working with the data:

3. *What data do you need to answer the question?*
4. *Where is the data coming from (identify all sources) and how will you get it?*
5. *Is the data that you collected representative of the problem to be solved?*
6. *What additional work is required to manipulate and work with the data?*

Deriving the answer:

7. *In what way can the data be visualized to get to the answer that is required?*
8. *Does the model used really answer the initial question or does it need to be adjusted?*
9. *Can you put the model into practice?*
10. *Can you get constructive feedback into answering the question?*

tools, at the right time, in the right order, to the address the right problem.

Congratulations! You passed!

Grade received 100% **Latest Submission Grade** 100% **To pass** 66% or higher

Retake the assignment in **7h 39m**

Go to next item

1. Feedback is not required once the model is deployed because the Model Evaluation stage would have assessed the model and made sure that it performed well.

1 / 1 point

- ○ True
- ◉ False

Correct
Correct.

2. A data scientist determines that building a recommender system is the solution for a particular business problem at hand. This is represented by the Modeling stage of the data science methodology?

1 / 1 point

- ○ True
- ◉ False

3. A car company asked a data scientist to determine what type of customers are more likely to purchase their vehicles. However, the data comes from several sources and is in a relatively "raw format". What kind of processing can the data scientist perform on the data to prepare it for the Modeling stage?

1 / 1 point

A. Feature Engineering.

B. Transforming the data into more useful variables.

C. Combining the data from the various sources.

D. Addressing missing invalid values.

- ○ Only options A and D are correct.
- ○ Only option C is correct.
- ○ None of the options are correct.
- ◉ All of the options are correct.

Correct
Correct.

4. Which of the following represent the two important characteristics of the data science methodology?

1 / 1 point

- ○ It immediately ends when the model is deployed because no feedback is required.
- ○ It is a highly iterative process and immediately ends when the model is deployed.
- ◉ It is a highly iterative process and it never ends.
- ○ It has no endpoint because data collection occurs before identifying the data requirements.

Correct
Correct.

5. For predictive models, a test set, which is similar to – but independent of – the training set, is used to determine how well the model predicts outcomes. This is an example of what step in the methodology?

1 / 1 point

- ◉ Model Evaluation.
- ○ Deployment.
- ○ Data Requirements.
- ○ Analytic Approach.

Correct

6. What are three important reasons that data scientists should maintain continuous communication with business sponsors throughout a project?

1 / 1 point

- [x] So that business sponsors can provide domain expertise.

Correct
Correct.

- [ ] Actually, data scientists do not need to maintain a continuous communication with business sponsors and stakeholders.
- [x] So that business sponsors can review intermediate findings.

Correct
Correct.

- [x] So that business sponsors can ensure the work remains on track to generate the intended solution.

Correct
Correct.

Lesson Summary

In this lesson, you have learned:

- The importance of stakeholder input.
- To consider the scale of deployment.
- The importance of incorporating feedback to refine the model.
- The refined model must be redeployed.
- This process should be repeated as often as necessary.

Mark as completed

Congratulations! You passed!

Grade received 100% **Latest Submission Grade** 100% **To pass** 66% or higher

Go to next item

1. The Modeling stage is followed by the Analytic Approach stage.

1 / 1 point

- ○ True
- ◉ False

Correct
Correct.

2. Select the correct statement(s) about the Model Evaluation stage of the data science methodology.

1 / 1 point

- ☐ Model Evaluation cannot include statistical significance testing.
- ☑ Model Evaluation includes ensuring the model is designed as intended.

Correct

2. Select the correct statement(s) about the Model Evaluation stage of the data science methodology.

1 / 1 point

- [ ] Model Evaluation cannot include statistical significance testing.
- [x] Model Evaluation includes ensuring the model is designed as intended.

Correct
Correct.

- [x] Model Evaluation includes ensuring that the model is working as intended.

Correct
Correct.

- [x] Model Evaluation includes ensuring that the data are properly handled and interpreted.

Correct
Correct.

3. The ROC curve is a useful diagnostic tool for determining the optimal classification model.

1 / 1 point

- (•) True

- [x] Model Evaluation includes ensuring that the model is working as intended.

Correct
Correct.

- [x] Model Evaluation includes ensuring that the data are properly handled and interpreted.

Correct
Correct.

3. The ROC curve is a useful diagnostic tool for determining the optimal classification model.

1 / 1 point

- (●) True
- ( ) False

Correct
Correct.

Back

Final Exam
Graded Quiz • 36 min

Due Aug 14, 11:59 PM IST

Congratulations! You passed!

Grade received 100% **Latest Submission Grade** 100% **To pass** 80% or higher

Go to next item

1. Select the correct statement.

1 / 1 point

- ○ The first stage of the data science methodology is Data Understanding.
- ○ The first stage of the data science methodology is Modeling.
- ◉ The first stage of the data science methodology is Business Understanding.
- ○ The first stage of the data science methodology is Data Collection.

Correct

2. What is an important stage in the data science methodology because it clearly defines the problem and the needs from a business perspective?

1 / 1 point

- ○ Data Collection

2. What is an important stage in the data science methodology because it clearly defines the problem and the needs from a business perspective?

1 / 1 point

- ○ Data Collection
- ○ Modeling
- ◉ Business Understanding
- ○ Data Understanding

Correct

3. According to the videos explaining the Data Requirements and Data Collection stages of the data science methodology, you can think of the Data Requirements and Data Collection stages as a cooking task, where the problem at hand is _____, and the data to answer the question is _____.

1 / 1 point

- ◉ The recipe; The ingredients
- ○ The shopping list; The store
- ○ The temperature; The shopping list
- ○ The cooking style; The appliance

Correct

4. In the Data Collection stage, techniques such as _____ and visualization can be applied to the data set, to assess the content, quality, and initial insights about the data.

1 / 1 point

- (●) Descriptive statistics
- ( ) The supervised method
- ( ) The unsupervised method
- ( ) Data manipulation

Correct

5. A _____ is used for predictive modeling.

1 / 1 point

- ( ) Technique set
- ( ) Modeling set
- (●) Training set
- ( ) Analysis set

Correct

6. A statistician calls a false-negative, a type I error, and a false-positive, a type II error.

1 / 1 point

- ○ True
- ◉ False

Correct

7. What stage encompasses all activities related to constructing the dataset?

1 / 1 point

- ○ The Modeling stage
- ○ The Data Requirements stage
- ◉ The Data Understanding stage
- ○ The Data Preparation stage

Correct

8. In what stage would you properly format the data?

1 / 1 point

- ○ The Data Requirements stage
- ◉ The Data Preparation stage
- ○ The Modeling stage
- ○ The Data Understanding stage

Correct

9. The final stages of the data science methodology are an iterative cycle between what stages?

1 / 1 point

- ○ Data Preparation, Evaluation, Feedback, Deployment
- ◉ Modeling, Evaluation, Deployment, and Feedback.
- ○ Data Understanding, Data Preparation, Evaluation, and Feedback.
- ○ Modeling, Deployment, Data Understanding, Data Preparation

Correct

10. Deploying a model into production represents the beginning of an iterative process from _____, then Model Refinement, and to Redeployment.

1 / 1 point

- ○ Scalability
- ○ Data Storage
- ◉ Feedback
- ○ None of the above

Correct

11. Select the correct sentence about the data science methodology as explained in the course.

1 / 1 point

- ○ The data science methodology does not depend on a specific set of technologies or tools.
- ○ The data science methodology always starts with Business Understanding.
- ○ The data science methodology is an iterative process.
- ◉ All of the above

Correct

- The data science methodology does not depend on a specific set of technologies or tools.
- The data science methodology always starts with Business Understanding.
- The data science methodology is an iterative process.
- (●) All of the above

Correct

12. What do data scientists typically use for exploratory analysis of data and to get acquainted with it?

1 / 1 point

- They use deep learning.
- They begin with regression, classification, or clustering.
- They use support vector machines and neural networks as feature extraction techniques.
- (●) They use descriptive statistics and data visualization techniques.

Correct